North Carolina
Statewide Technical Architecture

Data Domain

Office of the State Chief Information Officer
Enterprise Technology Strategies
PO Box 17209
Raleigh, North Carolina 27699-7209
http://www.ncsta.gov
ets@ncmail.net

Table of Contents

1. Principles:

1.1. Data is a valuable state asset.

Rationale:

- Data is a key State asset, managed by specified custodians on behalf of the citizens of the state. It is used to conduct business functions as required by law.
- The successful delivery of government services depends on conclusions derived from accurate, well-maintained, and secure data.
- Protecting the State's data is essential and must be performed subject to the laws, regulations, and policies governing data security and privacy. Failure to protect the State’s data greatly diminishes its value.
- As with any asset, data should be leveraged to maximize its potential.

1.2. Compliance with legal requirements in the design, implementation, and use of data by information systems protects the State from legal liabilities.

Rationale:

- Each Agency is responsible for knowing and following the Federal, State, and Local laws that apply to the data they maintain.
- Failure to comply with legal requirements exposes the State to legal liabilities.
- Systems and the data they contain may not be adequately protected when legal requirements are not met.
- Refer to authorized legal counsel for specific information regarding the legal requirements of storing and maintaining data that is owned by the State.

1.3. Implementaion of widely adopted standards for data design, definition, management, and exchange can facilitate the adapability, cost-effectiveness, and reuse of information.

Rationale:

- Use of existing, widely adopted standards leverages the expertise, resources, and vetting processes inherent in the standards development process.
- In the absence of specified Statewide architectural standards, use existing external standards when practical, as opposed to creating new standards.
- Examples of existing standards are as follows:
  - Industry Standards (e.g. ANSI, ISO, W3C, OASIS, IETF).
  - Federal Standards (e.g. NIST, FIPS).
  - De Facto Standards (e.g. ODBC, JDBC, ADO.Net).
  - NCSTA Standards.
  - Local standards.

1.4. Sound data management practices ensure the confidentiality, integrity, and availability of the State's data resources.

Rationale:

- Sound data management practices consist of establishing the appropriate processes, technology and personnel resources to ensure the confidentiality, integrity, and availability of the State's data resources.
- A high degree of data integrity can best be achieved through a statewide approach to data management.
- Sound data management includes:
  - Open, accessible, and adaptable database management systems (DBMSs).
  - Centralized data administration.
  - Highly granular and auditable security controls for access and authentication.
  - Robust backup and recovery capabilities and processes.
  - Replication facilities.
  - Database monitoring capabilities.
  - Data hygiene/cleansing capabilities.

1.5. Identification and use of authoritative sources of data within and across organizational boundaries enhances data accuracy and consistency.

Rationale:

- An authoritative source of data is trusted to be accurate and consistent. Authoritative data can either be created and defined within an agency or derived from external sources and maintained within an agency.
- While data can be used by more than one organization, a single organization must be identified as the authoritative source for that data. For example, "North Carolina Drivers License Number" is used by more than one agency, but DOT issues the drivers license and therefore is the authoritative source for this data. Non-authoritative data may be used by one or more state agencies, but a State agency is not identified as the authoritative source. For example "Social Security Number" is used in many applications and agencies across the state, but a State agency does not issues Social Security numbers.
- Each organization must identify and document the authoritative source for all data that it uses in an information system.
- Each data element should have a defined authoritative source and assigned custodian. The custodian is responsible for defining, establishing, and maintaining the accuracy of data collected into the authoritative source. Custodians must:
  - Provide accurate business definitions of data.
  - Develop enterprise-wide business views of shared data.
  - Provide business drivers to support centralized data administration.
  - Make complete metadata available for each data element shared from the authoritative source.
  - Define security requirements for data.

1.6. Operational and decision-support systems require high quality data.

Rationale:

- Information of appropriate quality is essential to making good decisions.
- All efforts should be made to ensure data is accurate, complete, current, and optimized for its intended use.
- The usefulness of the State's data depends on confidence in the quality of that data.

1.7. Reuse and sharing of data within and across organizational boundaries greatly reduces redundancy, cost, and leads to improved decision-making.

Rationale:

- Information systems must be designed to accommodate decision-making and data sharing beyond the borders of an organizational entity to address the larger communities of interest.
- Sharing of data greatly reduces data redundancy, entry, cost, and maintenance efforts.
- Consistent shared data definitions ensure data accuracy, integrity, and consistency.
- Data sharing reduces the overall resources required to maintain data across the enterprise.
- Data sharing should leverage an established infrastructure for widespread data access.
- For more information refer to the Application and Integration domains.

1.8. Data sharing among current and future systems and organizations is best accomplished when data is uniquely and accurately defined.

Rationale:

- Clear, concise, and unambiguous data definitions (metadata) are key to maximizing the value of data resources and resource investment. They aid in meeting customer needs, support the maintenance of the data resource through its lifecycle, and facilitate data sharing.
- Data meaning is clarified through data element definitions prepared in a standard manner from both a business and technology perspective. The business perspective establishes the semantic meaning of data. The technology perspective establishes the syntactical meaning of data.
- For more information refer to the Integration Domain.

1.9. Effective and accurate data exchange can best be achieved through formal processes and agreements.

Rationale:

- Information assets collected and maintained by an agent (e.g. an Agency) of the State are owned by the State. Agencies serve as custodians of data and must protect the data consistent with any legal requirements including privacy and confidentiality related legislation. Agencies must take the appropriate steps to protect the privacy and confidentiality of data in storage as well as when providing other Agencies with the data via formal exchange processes.

- Formal data sharing agreements must be in place prior to entities exchanging data that is protected by privacy, confidentiality, or other protective constraints.
- Exchanged data must be supported by sufficient documentation (semantic and syntactic meaning) to allow the receiving entity to determine the fitness for use for a particular application.
- Refer to authorized legal counsel for specific information regarding the legal requirements of exchanging data that is owned by the State.

1.10. Operational and analytical data require separate and distinct storage approaches.

Rationale:

- Online Transactional Processing (OLTP) applications perform mission critical day-to-day operations of the Enterprise at a detailed administrative level. Online Analytical Processing (OLAP) applications are used to for decision support purposes.
- Separating operational and analytical data maximizes the efficiency of both application types. Analytical data should be separate, distinct and extracted from operational systems.
- Large-scale separation of operational and analytical data is acceptable due to mutually exclusive design, processing and performance requirements. Separation enables the data to be optimized for specific purposes.
- OLAP and OLTP data should be stored in separate data stores.

2. Technical Topic: Data Sharing

2.1. Practices:

2.1.1. Utilize the State's data sharing services for data exchange.

Rationale:

- The custodian organization (Provider) of source data is responsible for developing and documenting any shared service(s) to access the data. This ensures data integrity and proper data interpretation.
- The organization (Consumer) requesting the data is responsible for developing the request to retrieve shared data according to the shared service specifications.
- The North Carolina Service Broker (NCSB) is a proven, cost-effective, enterprise-class service available to organizations that plan to share data.
- The State's interface engine can be used for sharing of legacy platform data or other data where the application source code cannot be modified or interfaced.

2.1.2. Use widely adopted standards for data exchange.

Rationale:

- If data needs to be shared across application or organizational boundaries then the data must be exchanged through exchange mechanisms as specified in widely adopted standards.

- The principle method for data sharing across application/organizational boundaries should be based on the current versions of the eXtensible Markup Language (XML) and Simple Object Access Protocol (SOAP) standards.
- Industry, international, and national standards bodies have established XML schemas to support a wide array of business information exchange requirements.
- Organizations should identify and adopt existing, industry-standard, XML schemas instead of creating new XML schemas.
- For more information refer to the Integration Domain.

2.1.3. Use non-proprietary formats to facilitate document exchange between users and applications.

Rationale:

- Proprietary formats inhibit document exchange, and may create barriers to communication.
- If proprietary formats are used, the capability must be provided to convert documents to standard formats for content exchange.
- Any vendor using proprietary formats should provide a conversion routine.
- Do not assume that the recipient of a document has the required application necessary to open, view and edit the document.
- Proprietary formats may accompany exchanged documents provided non-proprietary formats are also exchanged.

2.1.4. Use existing industry standard eXtensible Markup Language (XML) schemas for sharing data among disparate systems.

Rationale:

- The eXtensible Markup Language (XML) provides a critical foundation for enterprise data architectures. XML has matured as the Industry and Government standard for moving and sharing information both among different entities and systems, and even among components of a system. XML provides an opportunity for organizational entities to define and standardize XML schemas for their functions and for interactions with other internal and external entities such as Federal and Local Governments or Industry.
- Organizational entities should leverage existing industry standards such as e-Business XML (http://www.ebxml.org), JusticeXML (http://it.ojp.gov/jxdm; http://ets.state.nc.us/NCSTA/docs/Whitepapers/Justice%20XML%20Data%20Model%20Review.pdf) or join with Federal and Local Governments to define joint XML schemas that provide data interoperability across the tiers of government.
- All data centric development initiatives should define and implement an approach for using XML. Where new development or re-development efforts are pursued, XML should be considered for use as the default format for highly structured data as well as relatively less highly structured information.
- For legacy repositories that do not directly support XML, legacy to XML mapping and data transformation can be used to support interoperability across the data architecture.
- Development initiatives should work with communities in the relevant entities to define State-wide XML standards. Where possible, these standards should leverage

XML data elements and schemas that have been specified by voluntary consensus bodies as commercial and industrial standards.

- Use XML Schemas instead of XML Document Type Definitions (DTDs), which are now considered obsolete.
- W3C REC-XML- 20040204 (http://www.w3.org/TR/REC-xml/) establishes the standard for XML.

2.1.5. Systems that collect, utilize, or exchange geospatial information must comply with the policy statements and the standards adopted by the North Carolina Geographic Information Coordinating Council (GICC).

Rationale:

- The GICC is established by General Statute §143-725 and is responsible for improving the quality, access, cost-effectiveness, and utility of North Carolina's geographic information and promotes geographic information as a strategic resource for the State. The GICC creates policy, resolves technical issues, and approves standards related to North Carolina geographic information and GIS systems. (http://cgia.cgia.state.nc.us/cgia/)
- The North Carolina Center for Geographic Information and Analysis (CGIA) is staffed to the GICC. The CGIA promulgates standards approved by the GICC. CGIA has over 25 years of progressive GIS related experience and is nationally and internationally recognized as a subject matter expert in the field of GIS. The mission of the CGIA is to enhance, facilitate, and promote the efficient, cost-effective development and use of geographic information in North Carolina. CGIA provides enterprise class resources for sharing geospatial data among local, state, and federal bodies. (http://cgia.state.nc.us/cgia/).

2.1.6. Use NCOneMap for the collection and dissemination of geospatial data.

Rationale:

- Usage of the GIS data hosted on NC OneMap eliminates the need and expense of redundant data collection efforts.
- NC OneMap is a comprehensive statewide framework for the collection and dissemination of geospatial data. NC OneMap serves as the geospatial backbone for data collection and access in support of geographic-based decision making at the local and statewide levels, and in support of national priorities such as the Geospatial One-Stop E-Government Initiative (http://www.geo-one-stop.gov/).
- The North Carolina Center for Geographic Information and Analysis (CGIA) facilitated the development of NC OneMap (http://www.nconemap.net/) under the direction of the North Carolina Geographic Information Coordinating Council (GICC).
- Access to geospatial data will be provided according to open consensus standards and specifications. For example, the Open GIS Consortium (OGC) Web Map Service, Web Feature Service, and Web Coverage Service define methods for requesting data via the web in the geographic area of interest. (http://www.opengeospatial.org)

2.1.7. Document digital geospatial data in accordance with the Federal Geographic Data Committee's (FGDC) "Content Standard for Digital Geospatial Metadata, FGDC-STD-001-1998."

Rationale:

- The North Carolina Geographic Information Coordinating Council (GICC) has adopted the Federal Geographic Data Committee's (FGDC) "Content Standard for Digital Geospatial Metadata, FGDC-STD-001-1998" as the common set of terminology and definitions for the documentation of digital geospatial data (metadata).
- Digital geospatial data documentation (metadata) is necessary to provide data users with specific information about the lineage of a data set so it can be used with full knowledge of its source, quality, and content. Adequate metadata is required for the appropriate, responsible, and defensible use of any geographic data set.
- The FGDC Content Standard (http://www.fgdc.gov/metadata/contstan.html) establishes the names of data elements and compound elements (groups of data elements) to be used for these purposes, the definitions of these compound elements and data elements, and information about the values that are to be provided for the data elements.

2.1.8. Use mapping grade or survey grade Global Positioning System (GPS) systems for digital geospatial data collection and reporting efforts.

Rationale:

- The accuracy of data collected using GPS technology is highly variable, depending on field methods, type of equipment, and post-processing of the data. It is therefore imperative for North Carolina to maintain standards for GPS data collection, processing, and documentation.
- The quality of the GPS receiver also has a significant impact on positional accuracy. Survey grade receivers are the most expensive and produce data of the highest accuracy. Mapping grade receivers are mid-range in price and provide sufficient accuracy for most vocational and GIS applications. Data obtained using recreational grade receivers, the least expensive type of receiver, is the least accurate. Some recreational grade receivers have add-on modules that can be used for real-time differential correction. However, they cannot store data for post-processing. Recreational GPS receivers are not adequate for any state locational data acquisition.
- The North Carolina Geographic Information Coordinating Council (GICC) has established a standard for the accuracy of Global Positioning Systems (GPS) data. This standard requires all State government agencies to use either mapping grade or survey grade receivers in their GPS data collection efforts. ("Statewide Global Positioning System (GPS) Data Collection and Documentation Standards (http://www.cgia.state.nc.us/gicc/standards/gpstand.pdf).

2.2. Standards:

2.2.1. The standard file format for exchanging accessible, non-editable documents is Portable Document Format (PDF).

Rationale:

- PDF captures the elements of a printed document as an electronic image that can be viewed, navigated, printed, or sent via e-mail. PDF files preserve the original graphic appearance of documents.
- PDF files can only be created or edited using specialized software produced by Adobe (Adobe Acrobat)
- An Adobe Acrobat reader is necessary to view PDF files. This free reader is widely available on the Internet.
- Typical application software using this file format includes word processing, imaging systems, and World Wide Web publishing.
- PDF is a de facto standard.

2.2.2. The standard file format for exchanging facsimile and scanned documents is Tagged Image File Format for Image Technology (TIFF/IT).

Rationale:

- TIFF/IT is a common format for exchanging raster (bitmapped) images between application programs, including those used for scanning images.
- TIFF/IT files are used in desktop publishing, faxing, 3-D applications, and medical imaging. TIFF/IT files can be in any of several classes, including gray scale, color palette, or RGB full color.
- TIFF/IT files can include files with JPEG, LZW or ITU-T Group 4 standard run-length compression.
- ISO 12639:2004 specifies a media-independent means for electronic data exchange using a Tagged Image File Format (TIFF).

2.2.3. The standard file formats for exchanging raster based color documents, drawings, or photographs include Graphical Interface Format (GIF 89a) and Joint Photographic Experts Group (JPEG).

Rationale:

- GIF 89a provides a format for encoding relatively simple graphic images into bits for display on a computer screen. One advantage of this format is its ability to create an animated image, which is a single file that contains a set of images presented in a specified order. GIF 89a compresses the image (using the Lempel-Ziv Welch compression algorithm), making it easier to transmit and download. GIF 89a is typically used in the web development environment.
- GIF 89a is a de facto standard.
- JPEG is a graphical file format that provides digital compression and coding of continuous tone still images. JPEG images can be created in arrange of resolutions by specifying the image quality. Since the highest quality results in the largest file, trade-offs are made between image quality and file size. Most current applications that produce images in their respective proprietary formats provide an ability to save these images in JPEG format for non-editable data exchange.

- ISO/IEC 10918-4 establishes the standard for JPEG.

2.2.4. The standard file format for exchanging editable word processing documents is Rich Text Format (RTF).

Rationale:

- In most cases, parties exchanging editable word processing documents are using the same word processing application (e.g. Microsoft Word or Corel Wordperfect). However, in case where the parties in the exchange are not using the same application, RTF provides a means to exchange editable word processed documents and retain the majority of formatting applied to the source document.
- RTF is a desktop file format that permits exchange of text files between different word processing packages in different operating systems. For example, a file created in Microsoft Word 2000 in Windows XP can be saved as an RTF file that can be opened, read and edited by someone using WordPerfect 6.0 in Windows 98. RTF specifies details of the ASCII representation required for most low-level functions in word processing software. Information about fonts, page layout, and document management can be stored as part of the header information for each RTF file.
- RTF is a de facto standard supported by the majority of available word processors.

2.2.5. The standard file formats for exchanging computer aided design documents are the Initial Graphics Exchange Specification (IGES) and Data Exchange File (DXF).

Rationale:

- IGES enables Computer-Aided Design/Manufacturing (CAD/CAM) equipment to exchange product definition data throughout the life cycle of a given product. It allows digital exchange of product definition data in various forms (e.g., illustrations, two-dimensional drawings, three-dimensional edge-vertex models, surface models, solid models, and complete product models) independent of a particular CAD/CAM system.
- ANSI/US PRO/IPO 100 establishes the standard for IGES.
- DXF is a widely adopted exchange format used in Computer Aided Design applications (e.g. AutoCAD, Microstation) on small computer systems. It was developed primarily by AutoCAD developers and has thus received its popularity mainly from the high number of AutoCAD stations. Most CAD systems can export and most also import DXF, at least for two-dimensional data.
- DXF is a de facto standard.

2.2.6. The standard file format for exchanging moving images and audio is the Motion Picture Experts Group (MPEG-1).

Rationale:

- MPEG-1 provides for generic coding and compression of video and associated audio information. MPEG-1 defines techniques for compressing digital video by factors varying from 25:1 to 50:1.
- ISO/IEC 13818 establishes the standard for MPEG.

2.2.7. The standard file formats for exchanging editable vector based graphics such as line drawings are the Computer Graphics Metafile (CGM) and Scalable Vector Graphics (SVG) formats.

Rationale:

- The major advantage of vector graphics over the bit-mapped (raster) format is that vector graphics look the same, even when they are scaled to different sizes.
- Computer Graphics Metafile (CGM) is a long-standing data interchange standard that defines a neutral computer interpretable representation of two-dimensional (2-D) graphic information independent of any particular application or system. A CGM file can contain vector graphics, raster graphics and text. Most industry available graphics applications support export and import of CGM formatted graphics.
- ISO/IEC 8632:1999 establishes the standard for CGM.
- Scalable Vector Graphics (SVG) is the open XML standard developed by the W3C to describe 2D content that may contain vector graphics, raster images and font text. This new format brings the inherent benefits of vector graphics to the Web, combining them with all the capabilities offered by XML and related technologies. SVG is an entirely text based file format based on XML. Key advantages include smaller file sizes, indexable and searchable text elements, and high-resolution scan, zoom, and pan features.
- SVG is currently a Candidate Recommendation under the W3C (WD-SVG12-20030715).

3. Technical Topic: Data Design

3.1. Practices:

3.1.1. Design information systems that can accommodate rapid changes to data models and structures based on changes in business requirements or changes in database technologies.

Rationale:

- Business requirements change frequently. The data infrastructure and design must be adaptive and allow for changes to be easily implemented.
- Consider database scalability implications in the database design. Scalability allows for future growth, new database technologies and changes in business requirements.
- System designs must allow for potential replacement of the underlying database technology.

3.1.2. Design databases to be business driven and aligned with application services.

Rationale:

- Aligning data with application services facilitates changes in business processes. Only the data associated with a particular business process is potentially affected when a change is needed, not all the data associated with an entire application. It also increases performance for backup and recovery and provides higher reliability, availability, and scalability.

- In order to align data with application services, the following items need to be defined:
  - Business processes.
  - Data required to service the business processes.
  - Business units responsible for providing the business process.
  - Access to the data, and the know-how to use the data, by other business units or applications.

3.1.3. Design applications to ensure quality in the data capture process.

Rationale:

- Provide well-designed data-entry services that are easy to use (e.g., a GUI front-end with selection lists for standard data elements like text descriptions, product numbers, etc.).
- Validate data at every practical level to ensure data quality and avoid unnecessary network traffic. The system should be designed to reject invalid data elements and to assist the end user in correcting the entry.
- All reads and updates to an authoritative source database should occur using the business rules that own the data, not by direct access to the database.

3.1.4. Develop and document database designs from business, logical, and physical perspectives.

Rationale:

- There should be a direct mapping from the business view of the system design to the physical implementation of the database design at production.
- Use industry standard modeling conventions to capture the business view, business rules, and data flow of the application (e.g. Unified Modeling Language (UML)).
- Use Entity-Relationship (ER) Modeling to document the logical design of the database. The ER Model is a logical and graphical representation of the information needs of the application, without regard to the physical implementation or target technology. The ER Model should strive to achieve 3rd Normal Form.
- The Physical design evolves from the Logical design and is the specification of what is implemented in the target application. The Physical design should be optimized, efficient, buildable, and robust.

3.1.5. Provide configuration management for business, logical, and physical data designs.

Rationale:

- Database design models (business, logical, physical) store a wealth of system information and must be archived and protected.
- Configuration management principles must be applied to all database design models to facilitate maintenance, reengineering, and data sharing.
- Document, and maintain under configuration management, the "as built" lineage from the business view through the logical view to the physical implementation of the system.

3.1.6. Isolate, identify, and document the use of proprietary DBMS extensions that may create vendor lock-in.

Rationale:

- To differentiate their database products, many database vendors have implemented special extensions beyond the ANSI SQL-compliant features and command sets. Though these extensions may be useful for a particular function, usage must be limited.
- Use of proprietary extensions creates vendor "lock in" and should be avoided. In cases where use of proprietary extensions are required to meet business needs, then these extensions should be identified, isolated and clearly documented to aid in future application maintenance and portability.

3.1.7. Optimize the physical database design to maximize performance consistent with business requirements.

Rationale:

- As with all application and data access design, performance is always a factor to consider. However, database performance is only part of the total solution and must be evaluated in conjunction with other components that impact performance, such as network and application. When implementing data access, several practices can help optimize the physical implementation and operational use of systems, including:
  - Determine the appropriate amount of indexes in a database. When a record update occurs, not only the record is updated, but all the indexes are updated as well.
  - Limit the number of rows returned in a query. In OLTP, most users normally work with only a single row at a time or a few rows displayed in a grid, list or combo box. If a user will only be working with a small subset of records, there is no reason to return all the records in a table.
  - Return only the columns needed. Provide an explicit column list instead of a "SELECT ALL" query.
  - Limit the number of complex joins. Complex multi-table joins have negative performance ramifications.
  - Limit the rows used for pick lists, combo boxes, or lookup tables. If a large list is necessary, find an alternate method to provide the list.

3.1.8. Stored procedures and triggers must only be used for database structure maintenance.

Rationale:

- Stored procedures and triggers that support database structures, such as indexes, primary-foreign key relationships, date/time stamps and audit logging functions are permitted.
- Database access functions (e.g. Select, Insert, Update, and Delete) must be isolated in the application’s data access layer instead of in the database in the form of stored procedures and triggers.
- Use of stored procedures and triggers that tightly couple the application’s business rules to the database violates the principle of properly separating business rules and data access.

- For more information refer to the Application Domain.

3.1.9. Standardize on common design methodologies and tools for developing and maintaining the business, logical, and physical design of systems.

Rationale:

- Adoption of standard design methodologies and tools leverages the State's investment in tool purchases and analyst training.
- Standard methodologies such as Unified Modeling Language (UML), Business Process Design (IDEF0), and relational database design (IDEF1/X) are supported by the majority of enterprise level tools and are based on industry standards.
- Select design methodologies and tools that provide robust features for version control, configuration management, reverse engineering, metadata repositories, and XML data exchange.

3.1.10. Comply with ISO/IEC 11179 when defining data elements for custom designed solutions.

Rationale:

- ISO/IEC 11179, Specification and Standardization of Data Elements, is a widely adopted standard that: Describes standards for defining data elements to make them understandable and shareable.
- Provides guidelines for naming and defining of data elements.
- Provides information about the metadata that must be captured for data elements.
  - Following a standardized approach to defining data elements and capturing metadata (data about data) will facilitate application development, application maintenance, and sharing of data within and between applications

4. Technical Topic: Data Base Management System

4.1. Practices:

4.1.1. Use commercially-supported Relational Data Base Management Systems (RDBMS) for enterprise class applications.

Rationale:

- Relational databases offer dependability, flexibility, and compatibility for future data needs.
- Non-commercial open-source database products may not have undergone sufficient testing and lack organized support.
- Desktop database products (e.g. Microsoft Access, Microsoft SQL Server Desktop Engine) should not be used as the primary data store of business critical systems. However, they may be appropriate as part of mobile solution that allows synchronization between data maintained on an end-user platform (e.g. laptop, PDA, smart-phone) and a back-end server based application.
- Non-relational technology such as flat files can be used for temporary work storage and unstructured data such as textual data.

4.1.2. Separate the data sources for Online Transaction Processing (OLTP) data and Online Analytical Processing (OLAP) information.

Rationale:

- Separate data sources isolate OLTP systems, which perform mission critical business processing, from large ad hoc queries and online analytical data processing. If the data sources are not separate, ad hoc queries and direct access of data for OLAP systems can adversely impact online transactional processing.
- Data design is adapted for optimal performance for each type of application, OLTP or OLAP. For optimal performance, OLTP and OLAP may require different database designs. OLAP typically includes complex operations involving time series, dimensional, and trend analysis, which do not perform well with relational database technology alone (e.g., sometimes other methods of data storage are needed to support OLAP, such as multi-dimensional databases or flat files).

4.2. Standards:

4.2.1. Relational Data Base Management Systems (RDBMS) must meet ANSI SQL standards (e.g. SQL1999 and SQL 2003).

Rationale:

- Use of a RDBMS that meets ANSI SQL standards provides many safeguards including fully featured database services, protection against vendor lockin, and database portability.
- Vendors often implement extensions to standards to differentiate their products from their competitors. Use of vendor specific extensions will impact application portability should the underlying database need to be replaced.

5. Technical Topic: Data Access

5.1. Practices:

5.1.1. Data access objects must be properly seperated from business rules.

Rationale:

- This design facilitates database relocation, restructure, or platform changes with minimal disruption to the applications that use them. It is essential for adaptive systems.
- For more information refer to the Application Domain.

5.1.2. Data access must occur through an abstraction layer or service.

Rationale:

- There must be no direct access to data by the end-user or user interface. The client application should communicate with the database through data access objects by way of the business logic. This practice ensures security, data integrity, and accurate interpretation of the data and allows for adaptability to changes in business requirements.

- For more information refer to the Application Domain.

5.1.3. Design data access logic for reuse within and across applications.

Rationale:

- A typical application has numerous business rules. The data access logic for these business rules should be shared through a minimal number of reusable data access rules. Due to the commonality of database queries, many similar queries can execute using a single, properly planned data access routine.
- Portability to another database platform or vendor is simplified by having a fewer, well-defined data access rules.

5.1.4. Control and closely monitor ad hoc query access to OLTP and OLAP databases.

Rationale:

- Allowing end-user ad hoc access to OLTP and OLAP databases may circumvent critical business rules and negatively impact the performance of the database.
- Ad hoc queries on operational data may produce incorrect or misleading results due to the volatility of the underlying data.

5.1.5. Perform all data updates to authoritative sources of data, and replicate changes to remote databases as required.

Rationale:

- An authoritative source for data is the system where data is collected and maintained by the organization and/or application that is the custodian of that data. All other data stores must synchronize to the authoritative source. All data updates must occur against the authoritative source through the business logic and data access rules that encapsulate that data.
- Multi-tier, service-based application architectures facilitate the implementation of reusable business logic and data access rules.
- For more information refer to the Enterprise Application Integration Domain.

5.1.6. SQL must meet ANSI SQL standards (e.g. SQL2003 and SQL 1999 respectively).

Rationale:

- Use of proprietary SQL extensions, such as Microsoft's Transact-SQL or Oracle's PL/SQL, creates vendor lock-in and should be avoided.
- Vendor specific extensions to SQL in system development or via vendor supplied data access tools/protocols should be avoided. If vendor extensions cannot be avoided due to business or performance requirements, then the use of these extensions should be identified, isolated and clearly documented to support future system maintenance and portability.

5.2. Standards:

5.2.1. Use industry or de facto standards for database connectivity.

Rationale:

- Use of widely adopted industry standard approaches to database connectivity removes dependencies on the underlying database platform and tools.
- The goal of database connectivity is to make it possible to access any data from any application, regardless of which database management system (DBMS) is handling the data. Database connectivity takes a layered approach that inserts an Application Programming Interface (API) and database driver between an application and the DBMS. The purpose of this layer is to translate the application's data queries into commands that the DBMS understands. For this to work, both the application and the DBMS must be compliant with the database connectivity technology.
- Examples of database connectivity technologies include JDBC, JDO, JSQL, ADO.Net, OLE DB and ODBC.

6. Technical Topic: Data Management

6.1. Practices:

6.1.1. Centralize data storage to minimize data redundancy and facilitate access for shared and current data.

Rationale:

- High-volume transaction data that is shared across locations and that needs to be current for all locations must be stored centrally so all locations have access to the same data source.
- Data must be centralized when one or more of the following criteria occur:
  - Many users need access to latest data (i.e., OLTP systems).
  - The number of users is small and there are no distributed sites.
  - There is a lack of skills and tools at multiple sites to manage distributed data.

6.1.2. Implement policies, processes, methods, and tools to provide high-level data quality and consistency across distributed platforms.

Rationale:

- Both business users and Information Technology (IT) staff are responsible for data quality and consistency. Policies and procedures must be established to ensure the accuracy of data.
- IT staff are responsible for and must provide security mechanisms to safeguard all data under IT control. Business users must determine functional security requirements, while the physical security must be provided by IT.
- Applied systems management provides safeguards against data loss and corruption and provides the means of recovering data after system failures. This implies that effective backup and recovery systems are imperative and that data can be recovered in a timely basis regardless of the cause of loss.
- For critical functions, plan for business continuity under both normal and degraded operations.
- For more information refer to the Enterprise Management Domain.

6.1.3. When data replication is needed, evaluate the replication options and select the implementation that best meets the existing business requirements.

Rationale:

- Replication is the process in which files or databases are duplicated locally or remotely to facilitate business value and mitigate risk. Replication is used in areas such as business continuity planning, disaster recovery, and data warehousing.
- Replication is typically one-way, as when data is copied from a master server to one or more sets of replicas used for query-intensive applications. Replicas are used for reading, and potentially manipulating, data, but data is always created, updated and deleted at the master database.
- Replication solutions must be based on establishing functional and performance requirements such as source data volatility, topology, data refresh rates and schedules, processing overhead, support tool availability and capabilities, and network impact.

6.1.4. Use industry accepted standards, protocols, and practices for data synchronization.

Rationale:

- Data synchronization is a two-way process that involves capturing changes made to a remote copy (or subset) of a database (typically on a remote or mobile computing device) and propagating those changes back to a master database (or vice versa). As opposed to replication, where a copy of the master data is designed for read-only access, synchronized data is specifically designed to capture and propagate changes between the remote database and master databases.
- Data synchronization requires a means of maintaining a change record to the remote data.
- When data is updated in two or more places, there is always a chance of data integrity or accuracy conflict. Conflict resolution is the process, tools, and protocols required to detect and reconcile data consistency issues between source and target databases.
- Synchronization Markup Language (SyncML) provides a standard means of synchronizing data between remote and master databases. SynchML is an eXtensible Markup Language (XML) protocol under development as an open standard for the universal synchronization of data between devices. SyncML leverages existing standards such as MIME, vCard, and iCalendar. (http://www.syncml.org)

6.1.5. Plan and budget for the ongoing operations, administration, and maintenance of a operational and decision support systems.

Rationale:

- A support plan for information should be documented and implemented and considered as part of the Total Cost of Ownership of the system.
- The design of operational and decision support systems should be scaleable to meet future demands.

6.2. Standards:

6.2.1. Conform to the North Carolina Public Records Law (General Statute 132).

- Rationale:
- Specifically, any database that falls under the Public Records Law (North Carolina G.S. § 132-6.1 (b)) must be indexed according to the guidelines established by the N.C. Division of Archives and History and published as "Public Database Indexing Guidelines and Recommendations". A copy of this publication can be obtained by calling (919) 733-3540 or at http://www.ah.dcr.state.nc.us/e-records/default.htm.
- The North Carolina Government Information Locator Service (NC GILS) provides a standard protocol for Agencies to comply with the Public Records Law when indexing their databases. Through NC GILS, Agencies can index public databases consistently using widely known common elements and principles that ensure interoperability with other indexing systems. For more information about NC GILS, refer to the "Guidelines for Agencies Using the North Carolina Government Information Locator Service (NC GILS)" available at http://www.ncgils.state.nc.us/)
- For more information regarding the retention and destruction of any electronic records, refer to http://www.ah.dcr.state.nc.us/e-records/default.htm or contact the State Records Center staff at (919) 733-3540.

7. Technical Topic: Data Security

7.1. Practices:

7.1.1. Perform a risk assessment for the application database and data elements to determine level of security required.

Rationale:

- To assure adequate protection of data assets, perform a risk assessment to identify specific security concerns that must be addressed before development of an application.
- The security analysis will determine what measures must be put in place to restrict end users and applications from viewing, modifying, or deleting public and confidential data. The security analysis may reveal that adequate measures are in place to restrict end users and applications from viewing, modifying, and deleting low impact and public data.
- Classify users according to their functional data needs (e.g., outside access from business partners, citizens, suppliers, etc.).
- Authorize user access and update to data based on the principle of "least privilege".
- For more information refer to the Security Domain.

7.1.2. Use role-based access controls to streamline system administration and enhance system scalability.

Rationale:

- A role is a classification of users sharing the same security privileges.
- Granting permissions using roles, rather than assigning these permissions to each individual user, eases application development, management, and security.

Implementing role-based security improves application scalability since each process is not tied to a specific user.

7.1.3. Implement data security to allow for changes in technology and business requirements.

Rationale:

- Implement security to be a roadblock to unauthorized access, but not a hindrance to access by authorized users.
- Utilize enterprise level identity and access management services.
- Monitor ITS and industry security alerts and recommendations and implement changes to security configurations as required.
- For more information refer to the Security Domain.

7.1.4. Handle confidential data carefully.

Rationale:

- Confidential data must be secured to meet regulatory requirements, with proper policies and procedures in place to protect the data from unauthorized access.
- Confidential data must not be stored on a local desktop or laptop computers without password protection and/or encryption.

7.1.5. Audit user connections and transactions against database records.

Rationale:

- Record information about users and their connections as they update and delete data. The information that should be captured includes:
  - The user account used by the user to log into the system.
  - The TCP/IP address of the connected user's workstation.
  - The certificate information (if using certificates) about that user.
  - The old values that were stored in the record(s) before the modification.
  - The new values that were input to the record(s).

7.1.6. Implement and protect database transaction logs to facilitate the integrity of data

Rationale:

- Transaction logging records activity on the database and can be used to roll back a transaction.
- Protect the transaction log through access control and backup. Only the database should be writing to the transaction log. All other access should be read only.
- The transaction log should be located on a separate physical disk if possible. If not possible, use RAID technology to protect the integrity of the log file.

7.1.7. Implement security scanning, anti-viruses protections, file integrity software, and intrusion detection on the database server.

Rationale:

- Intrusion detection provides information about attempted attacks.

- Scan the database and database server for vulnerabilities. Take appropriate action to resolve and mitigate additional vulnerabilities.
- Monitor the database for possible intrusions. For example, monitor and alert when multiple invalid login attempts occur.
- Audit and review user logins, user account creation, and failed login attempts on a regular basis.
- For more information refer to the Security Domain.

7.1.8. Ensure data integrity by securing data movement or data transport.

Rationale:

- When high-impact, confidential data is transported through the LAN, WAN, or Internet, ensure that the data is encrypted and protected from alterations. This can be accomplished through Secured Socket Layers (SSL) or Virtual Private Network (VPN).
- If there is a risk of data being altered, then the data must be encrypted and protected.

7.1.9. Protect source code in data access objects, particularly if it contains password or other sensitive information.

Rationale:

- When using an multi-tier, service-based application development approach, it is common to encapsulate and abstract data access to work with business logic components. If an application needs to store account, password, or otherwise sensitive information then this source code must be protected from unauthorized access.
- Passwords must be stored in an encrypted or hashed format in accordance with state policies.
- For more information refer to Application Domain.

7.1.10. Do not store credit card numbers in the database. Store authorization numbers and discard credit card numbers after use.

Rationale:

- Storing credit card numbers and expiration dates in a database, even encrypted, can present an unjustifiable risk for the state.
- A credit card number is only necessary to request authorization. Keep the credit card number only until authorization is complete, then discard the card number and expiration date. The authorization number can be used to track activity and verify authorization.

7.1.11. Follow industry accepted practices for user identification, authentication and password management for database access. Change all default database passwords (i.e., SA accounts, etc.).

Rationale:

- Change all default database passwords (I.e. System Administrator (SA) accounts). SA accounts have full access to all databases in a database server. Hackers often attempt a login to a system administrator account using a default password. As soon as a database is set up, change all default passwords.

- For more information refer to the Security Domain.

8. Technical Topic: Data Warehouse

8.1. Practices:

8.1.1. Design data warehouse(s) to provide value to the State's business, and understandability to the end-user at an acceptable level of performance.

Rationale:

- A complex data warehouse that suffers from poor performance is a failure no matter how elegant the rest of the design may be because the people for whom it was intended won't want to use it.
- Complex business decision making requires rapid access to the high quality data.
- Analytical data is typically extracted, transformed and loaded into data warehouse(s) from disparate transactional systems.
- Information presented to data warehouse consumers should not overwhelm the user in its complexity.
- Design the data warehouse to optimize performance at the presentation, business logic, network and database tiers.

8.1.2. Develop data warehouse systems incrementally.

Rationale:

- Data warehouse efforts begin with a strategic set of information, preferably targeting a cross section of users.
- Critical Success Factors for developing data warehouses include:
  - Start small. Start by putting a relatively small amount of strategic data on a separate server, solving the problems of a specific set of users. The initial development effort should be time-boxed such that the first stage should be no more than 90 days from initiation to implementation.
  - Once the data is ready, add on-line analytical processing (OLAP) capability. For complex decision support, add on-line analytical processing, including trend analysis, indexing technology, and multi-dimensional database technology.
  - Implement near-real time production data feeds to maintain data freshness.
  - Apply data extraction technology to accomplish real time or near-real time feeds to maintain the data.

8.1.3. The data warehouse design process should identify specific requirements for data availability, freshness (i.e., live, 24 hours old, etc.), and recoverability.

Rationale:

- Some data warehouses need to be updated more frequently than others depending on business requirements. When the source data changes frequently, it may be necessary to update the data warehouse on a near-real time basis.
- If the source data is fairly stable, the data warehouse may only need daily, weekly, or even monthly updates. For example, a data warehouse that stores criminal data contains more volatile information and needs to be updated more frequently than a data warehouse that stores state registered corporation name and address information for public access.

8.1.4. During data warehouse design, determine the logic needed to convert the data, plan and generate the extraction and transformation routines, and assure the quality of the data populating the data warehouse.

Rationale:

- Data extraction and transformation is an important process for populating the data in a data warehouse and for ensuring that the data in a data warehouse is accurate.
- Planning for data extraction and transformation should start at the same time the data warehouse design starts.
- Data extraction and transformation logic includes data conversion requirements and the flow of data from the source operational database to the data warehouse. The data warehouse will contain data from disparate operational data sources. Data hygiene activities will be required to resolve any data inconsistencies before this data can be stored for analysis in the data warehouse.
- Develop a schedule for data extraction that both meets the needs of the data warehouse users and does not impact an OLTP system.
- Evaluate the impact of data extraction to any OLTP systems accessed.

8.1.5. Direct all information queries against decision support databases, not OLTP databases. Conversely, operational transactions should be directed to operational databases only, not OLAP databases.

Rationale:

- Permit only read-only access to end users of data warehouses.
- Data warehouses, and data marts, contain data that has been checked for consistency and integrity, and represents a cross-functional view of data.
- Data in transaction (OLTP) systems typically support a specific business group or function.
- OLTP transactions should not depend on a data warehouse database. They require a stable operational environment that is not affected by ad hoc usage or external data. Updates should only occur to the operational (OLTP) source where the data originates.

8.1.6. Assess the source data that will populate a data warehouse for quality and apply industry standard data hygiene practices to this data prior to importing into data warehouses.

Rationale:

- Data needs to be accurate to ensure good business decisions.
- Data needs to be relevant to the business need and consistent across multiple sources.
- Data must be complete. Data must contain the information necessary to answer the data warehouse business need.
- The data assessment also involves evaluating the business rules associated with that data. The appropriate business rules must be applied to the data to maintain accuracy.

8.1.7. Perform periodic validity audits against the data warehouse information model to ensure a high level of confidence in the quality and integrity of the data.

Rationale:

- Accelerated decision-making requires high quality data. If operational data has changed or additional data is needed, changes must be made in the information model and in the data warehouse itself.
- The data stored in a data warehouse should conform to the information model.
- The source data populating a data warehouse should be verified for consistency and accuracy.
- The data warehouse should still correspond to business needs.
- Ensuring the integrity and quality of data is the responsibility of both the business users and Information Technology staff.